Bonnie Turek

Eco 602 – Week 3 Reading Questions

9/13/2021

Q1. Which of below plots show every single data point?

* Histogram – NO, these show bins, or categories of data observations
* **Scatterplot** – YES, these show all observations between two variables
* **Cleveland dotplot** – YES, useful to identify outliers and homogeneity for all observations which are represented as single dots
* Boxplot – NO, these are graphically represented by median, quantiles, and range
* **QQ plot** – YES, every observation shown but just with rank and z-standardizations performed (i.e. comparing distribution of sample pts. to theoretical normal distribution)
* **Coplot** – YES, similar to scatterplot, all observations between two variables are shown, but with a third conditional element which breaks the scatterplots up from bottom left to top right

Q2. Which of the plot types show aggregated or summarized data?

* **Histogram** – YES, summarized to show center and distribution of data
* Scatterplot – NO, shows raw values of all data points
* Cleveland dotplot – NO, typically not aggregated, they show every observation as a dot
* **Boxplot** – YES, shows a summary of the spread of a variable and it’s related mean and quantile values
* QQ plot – NO, this data is not aggregated, just transformed via standardizations
* **Coplot** – YES, although all of the data points are shown, I guess you could consider the data to be aggregated since you are breaking them into categories based on a third variable

Q3. Conditional plot, conditioning variable, and related terms occurred throughout the Zuur and McGarigal readings. Explain what a conditional variable means in the context of graphical data exploration.

A conditional variable may be either a variable that takes on only a few discrete values or a continuous variable that is divided into a limited number of subsets. The conditional/conditioning variable is a way to subset a data sample and a comparison of one or more variables. For example, we can look at just the relationship between X and Y. But if we add a conditional variable, then we could look at the relationship between X and Y, given Z. A certain condition or circumstance (Z) must be met to create that subset of your data (X vs. Y). An example of a conditional variable in terms of graphical data exploration: you could be comparing the relationship between air temperature and annual precipitation (no conditional variable), or if you add a conditional variable, then you can examine air temperature vs. annual precipitation, but only at a certain elevation or higher (i.e. 100ft above MSL and up). Adding that conditional variable might subset your data into a smaller group, but it could enhance your data exploration, add value to the results and give you more insight about relationships and interactions among more than two variables in your dataset. A basic scatterplot is an example of one variable, X and it’s relationship to another variable Y plotted along two axes. However, if you add a third conditioning variable, Z, you can plot your X vs. Y variables in a Coplot (or conditional plot) to show more variability based on a third variable.

Q4. List *at least three* of the common measures of spread or dispersion that were mentioned in the readings:

Variance, Standard Deviation, Coefficient of Variation, Range, Interquartile Range

Q5. Choose *two of the measures* in your list and explain how they capture different aspects of the concept of spread:

Standard Deviation – This value measures the spread of a data distribution. The more spread out a data distribution is, the greater its standard deviation. In other words, the standard deviation is the root mean (squared) deviation from the mean or expected value. We get the standard deviation by taking the square root of the variance. As you add 1,2, 3 and more standard deviations from the mean (+ or -), you are capturing more and more of your data values. The standard deviation is especially important with a normal distribution. For example, 1 standard deviation away from the mean (+ or -) captures 68% of the values. 2 standard deviations captures 95% of the values and 3 standard deviations captures 99%. A standard deviation close to 0 indicates that the data points tend to be close to the mean.

Range – This is pretty simple to understand. The range is the difference between the smallest value and the largest value in a dataset. You would be referring to the range of a single variable. For example, the range of X would be the minimum value of variable X – maximum variable of X. Examining the range just allows you to examine where the extremes (min and max values) of a dataset may lie. This is different than the standard deviation, which tells you more about the distribution of data around the mean (or average expected value) of your dataset.

Q6. Consider a dataset that you have collected or worked with. List two of the important reasons to perform data exploration (numerical and/or graphical). For each of the two reasons you identify, describe the quantities or plots you would use and the insight you would gain:

* 1. Need to assign meaning to thousands of rows and columns of data points

When working with such a large dataset of sediment organic carbon data, for example, I would start by looking at the range and mean values to get an idea of how widespread of values your data is comprised of and whether it might be more manageable to break the large dataset up. I would also then create a box and whisker plot to help visualize the mean, interquartile range, and any outliers. It is specifically important to look at outliers because they may signal that there was perhaps a mistake in data processing or analysis or simply even a human typing/writing error. You would obviously want to know if outliers exist in the data, and examine them, so you don’t include any false or mistakenly labeled data points.

* 1. Visualize data in order to make initial interpretations and potentially frame new research questions

When working with predicting carbon content of sediment in marshes (“blue carbon”), carbon content can be a function of one or more variables combined. For example, carbon content can be related to elevation, distance to tidal creeks, distance to the marsh edge or creek outlet, or vegetation structure. It is certainly important for data exploration in order to visualize those initial comparisons between data collected in the field and via GIS tools. In this case, I would simply set up scatterplots or a scatterplot matrix of all of these variables each separately compared to corresponding carbon content value collected at those locations. This may help me get a better idea of the best predictor variables of carbon content. Or if there are no single variables that alone can explain carbon content, I can at least formulate more questions to guide future data collection or move forward and try creating a model that focuses on the interactions of 2 or more variables.

I did not work with other students on these reading questions.